



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) **EP 1 316 019 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
08.03.2006 Bulletin 2006/10

(21) Application number: **01968979.3**

(22) Date of filing: **30.08.2001**

(51) Int Cl.:
G06F 12/08^(2006.01) G06F 12/12^(2006.01)

(86) International application number:
PCT/US2001/041965

(87) International publication number:
WO 2002/019115 (07.03.2002 Gazette 2002/10)

(54) **MANAGING REPLACEMENT OF DATA IN A CACHE ON A NODE BASED ON CACHES OF OTHER NODES**

VERWALTUNG DER ERSETZUNG VON DATEN IN EINEM HILFSPEICHER AUF EINEM KNOTEN
AUFGRUND DER HILFSPEICHERN ANDERER KNOTEN

GESTION DU REMPLACEMENT DES DONNEES EN ANTEMEMOIRE DANS UN NOEUD SUR LA
BASE DES ANTEMEMOIRES DES AUTRES NOEUDS

(84) Designated Contracting States:
**AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE TR**

(30) Priority: **31.08.2000 US 652384**

(43) Date of publication of application:
04.06.2003 Bulletin 2003/23

(73) Proprietor: **Oracle International Corporation**
Redwood Shores, CA 94066 (US)

(72) Inventors:
• **PRUSCINO, ANGELO**
Los Altos, CA 94022 (US)
• **BAMFORD, Roger**
Woodside, CA 94062 (US)
• **BRIDGE, William, Jr.**
Alameda, CA 94502 (US)
• **BROWER, David**
Alamo, CA 94507 (US)

• **MACNAUGHTON, Neil**
Los Gatos, CA 95031 (US)
• **WANG, Tak, Fung**
Redwood Shores, CA 94065 (US)

(74) Representative: **Viering, Jentschura & Partner**
Steinsdorfstrasse 6
80538 München (DE)

(56) References cited:
US-A- 5 297 269 US-A- 5 680 576
US-A- 5 829 032 US-A- 6 092 157

• **FEELEY M J ET AL: "IMPLEMENTING GLOBAL
MEMORY MANAGEMENT IN A WORKSTATION
CLUSTER" OPERATING SYSTEMS REVIEW
(SIGOPS), ACM HEADQUARTER. NEW YORK,
US, vol. 29, no. 5, 1 December 1995 (1995-12-01),
pages 201-212, XP000584826**

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

EP 1 316 019 B1

Description

FIELD OF THE INVENTION

[0001] The present invention relates to cache management and, more particularly, to managing replacement of data in a cache on a node based on caches of other nodes.

BACKGROUND OF THE INVENTION

[0002] A cluster is a collection of nodes, each with a volatile dynamic memory devices ("dynamic memory") that is shared by one or more CPUs. Each node in a cluster is connected to the other nodes in the cluster via an interconnect. Typically, each node in the cluster shares resources with other nodes in the cluster. Such resources may be, for example, data items, such as pages stored on a static memory device (e.g. disk storage devices).

[0003] Software, referred to as an application, runs on each of the nodes. The application operates upon data from shared data resources. For example, a database system may run on each node in the cluster and manage data stored on data blocks on the disk storage devices.

[0004] To speed access to data items needed by an application running on a node in a cluster, copies of the data items are kept in the cache of the node. A cache is a storage medium used to store a copy of a data item for efficient access, where that data item may be obtained less efficiently from another source. The other, less-efficiently-accessed source is herein referred to as a secondary data resource. For example, a database system running on a node in a cluster typically stores data blocks durably on disk. However, the database system loads copies of the data blocks in a cache in dynamic memory to access them. Similarly, a browser running on a personal computer stores frequently accessed web pages, obtained over a network, in a cache in the form of disk files.

[0005] A cache on a node is managed by a cache manager. A cache manager is a system of software that attempts to manage a cache in a manner that reduces input and output between the cache and a secondary data resource. Typically, the cache manager keeps a copy of the most frequently accessed data items in the cache. In addition, a cache manager may track modifications made by an application to a data item, and may arrange, through "write back", "write through", and various logging mechanisms, that each data item is durably stored in a secondary data resource (e.g. disk storage device).

[0006] Data items frequently accessed by a particular application are collectively referred to as the working set of the application. When the working set of an application is too big to fit into the cache on a node on which the application is running, the application thrashes. Thrashing involves replacing cached data items in a cache with data items from secondary data resources at an unde-

sirably high frequency. Thrashing occurs when frequently accessed data items continuously replace each other in the cache, causing frequent cache misses and thereby drastically increasing the average data item access time.

When thrashing is occurring and an application needs data items, the application is too often delayed by waiting for the retrieval of a needed data item from the secondary data resource. Thrashing thus reduces throughput. On the other hand, when the working set can fit into available dynamic memory, the application does not thrash, and the throughput of the system is improved.

[0007] A scalable cluster is a cluster to which resources may be added to proportionally increase the capacity of the cluster as a whole. A resource may be, for example, a computer, a CPU, a storage device, interconnect bandwidth, or dynamic memory per computer. Typically, capacity is measured according to some measure of business growth, such as the number of users, web page hits per second, transactions per period of time, and/or terabytes of data.

[0008] As the need for capacity grows in a cluster, usually the need for CPUs grows, and sometimes the working set grows. To accommodate the need for additional CPUs, more nodes may be added to the cluster. However, accommodating an increased working set requires more than just adding computers to the cluster. Rather, the dynamic memory of all the nodes in the cluster must be increased because each node needs enough dynamic memory to store the working set. Each node needs enough dynamic memory to store the entire working set because a large portion of the frequently accessed pages are typically duplicated in the cache of each node.

[0009] For any computer, the amount of dynamic memory that may be added is limited. Even before the limit is reached, the total amount of dynamic memory required by the application is equal to its working set multiplied by the number of nodes in the cluster. As the cluster grows, if the working set grows as well, then the price-performance decreases with cluster size, and so the cluster does not scale. If the working set is larger than the limit, then the application thrashes. Adding nodes to the cluster may increase the CPU capacity of the cluster, but will not eliminate the thrashing. Consequently, the throughput of the cluster suffers.

[0010] Based on the foregoing, it is desirable to provide a mechanism to reduce thrashing on a cluster that does not require adding dynamic memory to each node in the cluster.

[0011] The following references are indicative of the present state of the art:

U.S. Pat. No. 6,092,1 S7 to Suzuki describes an information processing system using a plurality of nodes, in which each of the nodes is capable of accessing its own internal memory, as well as memory of the other nodes;

Feeley M. J. et al. "Implementing Global Memory Management in a Workstation Cluster Operating

Systems Review (SIGOPS), ACM Headquarter, New York, U.S. vol. 29, no. 5, 1 December 1995, pages 201-212, XP000584826 describing the design of a workspace cluster with the objective of using a single, unified but distributed memory management algorithm at the lowest level of the operating system; U.S. Pat. No. 5,680,576 to Laudon describes multiprocessor system having a plurality of requestors, a memory and memory directory controller employing directory-based coherence in system implements a method to detect dropping of clean-exclusive data. Only one intervention message is permitted to target an exclusive object held by a first requestor, wherein the intervention message is caused by a second requestor. The system detects whether the first requestor has an outstanding writeback for the object targeted by the intervention message, as well as whether the first requestor has a clean-exclusive, dirty-exclusive or invalid copy of the object targeted by the intervention message. A clean-exclusive copy of the object has been dropped when no outstanding writeback is detected and the first requestor has the object in the invalid state; U.S. Pat. No. 5,297,269 to Howard et al. describes a cache coherency protocol for a multiprocessor system which provides for read/write, read-only and transitional data states and for an indication of these states to be stored in a memory directory in main memory; and U.S. Pat. No. 5,829,032 to Yamagami et al. describes whereby as much tag information as corresponds to the number of blocks stored in main memories is stored in tag memories. The tag information indicates whether or not a processor belonging to a node other than the nodes containing tag memories has made an access request and the contents of a cache have been rewritten. Bus bridges perform cache coherency control, referring to the tag information. When the tag information indicates "modified," the bus bridges stop the data read from the main memories and send the correct data to the requesting processor after a copy-back process of the modified block has been completed.

SUMMARY OF THE INVENTION

[0012] The invention provides a method for managing the replacement of data stored in buffers of a plurality of caches, the plurality of caches residing on a node of a set of one or more nodes. Each node of the set of one or more nodes includes one or more caches of the plurality of caches, whereby the plurality of caches includes a first cache. The method includes selecting for replacement a first buffer from said first cache based upon one or more factors from a plurality of buffers available for replacement in the first cache. The one or more factors in a particular embodiment includes either the state of at least one other cache of the plurality of caches, or the

configuration of the at least one other cache of the plurality of caches. The method further includes the operation of replacing the first data item stored in the first buffer with a second data item, and transferring the first data item to another particular cache of the plurality of caches. Responsive to transferring the first data item to the particular cache, the method further includes generating data that indicates the first buffer contains data that may be removed from said first cache, whereby the one or more factors include data indicating that the first buffer contains data which is replaceable.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

FIG. 1 is a block diagram of an exemplary cluster; FIG. 2 is a flowchart of a process that implements techniques for managing a cache according to an embodiment of the invention; FIG. 3 is a flowchart of a process that implements techniques for managing a cache according to an embodiment of the invention; and FIG. 4 is a block diagram of a computer system upon which an embodiment of the present invention may be implemented.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0014] A method and apparatus for managing caches on the nodes of a cluster is described. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

EXEMPLARY CLUSTER

[0015] FIG. 1 is a block diagram that depicts an exemplary cluster 100 upon which an embodiment of the present invention may be implemented. Cluster 100 includes nodes 110, 130, and 150. Interconnect 160 connects nodes 110, 130, and 150, such that any node in cluster 100 may transfer data to the dynamic memory of another computer.

[0016] Each of nodes 110, 130, and 150 include a cache and a cache manager. Specifically, node 110 includes cache 114 and cache manager 112, node 130 includes cache 134 and cache manager 132, and node 150 includes cache 154 and cache manager 152. While

an embodiment of present invention is illustrated using caches that each reside on different nodes, the present invention is not limited to caches that each reside on different nodes. Rather, the present invention may be implemented with caches that include a subset of caches that reside on the same node, or that all reside on the same node.

[0017] Caches 114, 134, and 154 are managed respectively by cache managers 112, 132 and 152 to store data items accessed by a set of applications running on cluster 100. In an embodiment of the present invention, the set of applications may be a database system. Each of cache managers 112, 132, and 152 run under the control of a database system. Caches 114, 134, and 154 are used to store data blocks that contain data managed by the database system.

[0018] Cache managers 112, 132, and 152 collectively form a "coherent cluster cache manager". A coherent cluster cache manager is a system of software running on a cluster that globally manages the caches in such a way that when an application requests access to a data item in a given cache, and the data item is not stored in the given cache but is stored in another cache, a copy of the data item is copied directly from the other cache to the given cache. If the data item has been modified, a copy of the data item as modified may be transferred to the given cache without first durably storing the data item to a secondary data resource. An example of a coherent cache manager is disclosed in Klots, which describes a database system that manages a cache.

[0019] A coherent cluster cache manager differs from the typical conventional cache manager on a cluster. A conventional cache manager typically operates on a node in a cluster, and when the node requires a data item that is not cached on the node, it gets a copy of a data item from a secondary data resource, not from the cache of another node. In fact, the conventional cache manager has no knowledge that a copy of the requested data item exists in another cache, whether the cache resides on the same node or a different node. Thus, to get a copy of the data item, a conventional manager incurs the cost of accessing a secondary data resource, such as a disk storage system.

[0020] Accessing a secondary data resource, such as a disk storage system, entails relatively expensive input and output operations. On the other hand, accessing the dynamic memory on another node in the cluster is a relatively inexpensive operation. Thus, getting a copy of a data item from another cache, whether or not the cache is on another node, is far cheaper than getting a copy from a secondary data resource.

[0021] With the ability to more cheaply get a requested data item from another cache, duplicates of frequently accessed data items can be removed from some caches without causing thrashing, as long as they are retained in at least one cache in the cluster. Rather than re-retrieving a copy of a frequently accessed page from a secondary data resource, a copy is re-retrieved far less ex-

pensively from the cache of another node.

[0022] Because duplication of data items may be reduced without causing thrashing, apportioning the working set among the nodes in a cluster becomes feasible. Because the working set is divided between rather than duplicated on all the nodes, additional dynamic memory for a working set may be provided by adding a node to the computer, rather than increasing the dynamic memory of all the nodes. For the purposes of efficiency, the size of the working set is therefore not restricted by the dynamic memory capacity of a computer, and consequently, scalability is not limited by the size of the working set and the dynamic memory capacity of the nodes in the cluster.

[0023] For example, there are 1000 data blocks frequently accessed by a database system running on a four node cluster. When the working set is apportioned, the frequently accessed blocks do not have to be duplicated in the caches of the nodes. The working set may therefore be divided so that each of the caches holds 250 of the frequently accessed data blocks. If another node is added, then each of the nodes would have to hold 200 blocks. If the working set is not divided, then each node must have enough dynamic memory to hold 1000 blocks.

REPLACEMENT POLICIES THAT REDUCE DUPLICATION OF DATA ITEMS

[0024] The techniques described herein for reducing the duplicates of data items stored in global cache are incorporated into the replacement policy of a cache manager. The term global cache is used herein to refer to a set of caches of one or more nodes, for example, the nodes in a cluster, where each of the caches holds data items accessed by a set of one or more applications. A particular cache managed by a cache manager is referred to as a local cache with respect to the cache manager. A replacement policy is the set of rules and procedures followed by a cache manager to select a buffer that already contains a data item when a buffer is needed to hold a given data item. The data item in the selected buffer is replaced with the given data item. Consequently, the cache manager selects a buffer that contains data that has a relatively low cache retention value.

[0025] Cache retention value is a measure of the degree to which retaining a data item in the cache benefits the efficient operation of the cache. Typically, the data items that will be most frequently accessed have greater cache retention value because retaining them in cache avoids a greater number of input and output operations between the cache and secondary data source. The term retain refers to executing operations that increase the likelihood that a particular data item will remain in a cache longer than another data item, typically a data item with lower cache retention value. Retain may refer to selecting for replacement another data item, or simply assigning a lower cache retention value for the other data item.

[0026] A variety of factors may be used to determine

the cache retention value of a buffer. For example, the most recently accessed data items usually have a greater cache retention value than the least recently accessed because the most recently accessed data items are the most likely to be accessed in the future. A cache manager may maintain a least recently used list (LRU), which orders buffers from the most recently accessed to the least recently accessed. The most recently accessed items are at the top of the list and the least recently used are at the bottom of the list. Thus, the position of a buffer within an LRU indicates the cache retention value of the data item held within a buffer. Another factor that may be considered is how frequently a data item is accessed within a recent period of time. The data items most frequently accessed in the immediate past have a greater cache retention value because these data items are the most likely to be accessed in the near future. In determining the cache retention value according to the techniques described herein, a cache manager follows a replacement policy that also accounts for factors involving the global cache, not just one local cache. One of these factors is whether a data item is duplicated in the caches of a global cache.

REDUCING DUPLICATE COPIES OF A DATA ITEM

[0027] FIG. 2 is a flow chart depicting a process that may be followed to implement, at least in part, a replacement policy that reduces duplicate copies of data items in the caches of a cluster. The steps are performed by a cache manager when it receives a request for a data item in the cache being managed by the cache manager. The steps are illustrated using Cluster 100 shown in FIG. 1.

[0028] Referring to FIG. 1, at step 210, cache manager 112 receives a request from cache manager 132 to directly transfer a data item stored in cache 114 to cache 134. At step 220, the cache manager 112 directly transfers the requested data item to cache 134.

[0029] At step 230, it is determined whether the data item transferred is a data item that merits duplication in other caches. Generally, data items that merit duplication have such a high cache retention value that the cost of duplicating them in terms of, for example, increased dynamic memory usage, justifies their duplication. For example, the most frequently accessed data items in the cache 110 may merit duplication. If it is determined that the data item transferred is a data item that merits duplication, the steps shown in FIG. 1 end. Otherwise, execution of the steps proceeds to step 240.

[0030] At step 240, cache manager 112 establishes the buffer holding the data item as replaceable. Establishing the buffer as replaceable qualifies it for selection during replacement processing. There are a variety of ways to establish a buffer replaceable. For example, a buffer may be marked as being immediately replaceable by setting a flag in the buffer, or added to list of buffers that are replaceable.

[0031] When cache manager 112 performs replace-

ment processing, it may select a buffer that has been established as immediately replaceable. The buffer may be selected even though, for example, its place on an LRU maintained by cache manager 112 indicates that the buffer holds a data item whose recency of access is relatively high. When a cache manager selects for replacement a buffer that has been established as immediately replaceable because the buffer's data item was transferred to another cache, the cache manager is following a replacement policy that accounts for the state of global cache. Specifically, the cache manager is replacing the data item at least part because it is stored in another cache.

[0032] On the other hand, the existence of a copy of data item on another cache is but one factor that may be used to select a buffer for replacement of its contained data. Other factors to consider include those mentioned already, such as the position of a buffer in an LRU list.

ACCOUNTING FOR THE CACHE RETENTION VALUE BETWEEN DATA ITEMS IN DIFFERENT CACHES

[0033] A replacement policy that accounts for whether a data item has been duplicated on another cache is but one example of how the state of caches on multiple nodes in a cluster may be accounted for. FIG. 3 depicts another process that accounts for the state of caches on multiple nodes. The process is performed when one cache manager requests a data item stored in the cache of another node.

[0034] Referring to FIG. 3, at step 310, cache manager 112 receives a request from cache manager 132 to directly transfer a data item stored in cache 114.

[0035] At step 320, the cache manager 112 directly transfers to cache manager 132 the requested data item and the data that indicates the lowest cache retention value associated with any buffer in the cache 114. The buffer in cache 114 with the lowest cache retention value need not be the buffer in cache 114 in which the requested data item resides. Various mechanisms may be used to establish the cache retention values of buffers. For example, data that indicates the lowest cache retention value may be data that indicates the last time a data item was accessed, or data that specifies the frequency with which it was accessed within a period of time.

[0036] At step 330, the cache manager 132 receives the requested data item and the data indicating the lowest cache retention value ("received cache retention value").

[0037] At step 335, cache manager 132 selects a buffer in cache 134 for replacement with the requested data item transmitted from cache manager 112 to cache manager 132.

[0038] At step 340, the cache manager 132 determines whether the received cache retention value is greater than the cache retention value of the data item in the buffer selected for replacement (the "selected data item"). If selected data item has a cache retention value that is greater than the received retention value, then

control proceeds to step 350, where cache manager 132 transmits the selected data item to cache manager 112.

[0039] If, on the other hand, the received cache retention value is greater than the cache retention value of the selected data item, then at step 375, the requesting node replaces the selected data item with the received data item, without having transmitted the selected data item to the cache manager 112.

[0040] At step 350, cache manager 132 transmits the selected data item to cache manager 112. At step 360, cache manager 112 receives the selected data item. At step 370, cache manager 112 replaces with the selected data item a data item in a buffer with the lowest cache retention value in cache 114. Control then proceeds to step 375.

REPLACEMENT POLICIES BASED ON THE CONFIGURATION OF THE GLOBAL CACHE

[0041] The techniques described herein incorporate replacement policies that account for the global state of the caches in a cluster. Replacement policies may, however, account for global considerations other than the global state of cache. For example, the replacement policy for a particular cache may be based on the configuration of other caches. The configuration of a cache is the logic and values that govern the operation of cache that do not change during the operation of a cache. The state of a cache, on the other hand, is collectively the current values that change during the operation of the cache. For example, the state of a cache includes the data item in a buffer of a cache. In addition, the state of a cache includes values that represent an attribute of the buffer or the data item in the buffer that change as the cache is operated upon, such as the position of a buffer in an LRU, or the frequency of access of a buffer. The configuration of the cache may include software that governs the operation of the cache, and operational parameter values such as the size of the cache, or the number of buffers in the cache.

[0042] The following example illustrates how a replacement policy of a cache may be based on the configuration of other caches. Assume that a database system is running on a cluster, and that the caches on the nodes of the cluster are configured so that a cache is a "home" to a particular set of data blocks. A cache that is "home" to a particular set of data blocks is herein referred to as the primary cache for those set of data blocks. When a cache manager must select a buffer in cache for replacement, a cache manager first selects buffers with data items whose primary cache is another cache, before selecting data items without a primary cache or whose primary cache is the current cache. When a cache manager selects a buffer in the cache managed by the cache manager because the configuration of another cache makes it the primary cache of the data item in the buffer, a cache manager is following a replacement policy that accounts for the configuration of the another cache.

PRIMARY CACHES FOR SHARED NOTHING SYSTEMS

[0043] A further example of a replacement policy based on the primary cache of a data item is a replacement policy for a global cache where a primary cache for a particular data item is a cache on a node that may directly access the secondary data resource for a data item. Treating the cache as a primary cache for a data item in this manner may be particularly advantageous for a shared nothing system. A description of shared nothing systems is therefore useful.

[0044] There are three general types of multi-processing systems: shared everything, shared disk systems, and shared nothing systems. In shared everything systems, processes on all processors have direct access to all dynamic memory devices and to all static memory devices in the system. A high degree of wiring between the various computer components is required to provide shared everything functionality.

[0045] In shared disk systems, processors and dynamic memory devices are grouped into nodes. Each node in a shared disk system may itself constitute a shared everything system that includes multiple processors and multiple memories. Processes on all processors can access all static memory devices in the system, but only the processes on processors that belong to a particular node can directly access the dynamic memory devices within the particular node. Shared disk systems generally require less wiring than shared everything systems.

[0046] In shared nothing systems, all processors, dynamic memory devices and static memory devices are grouped into nodes. In shared nothing systems as in shared disk systems, each node may itself constitute a shared everything system or a shared disk system. Only the processes running on a particular node can directly access the dynamic memory devices and static memory devices within the particular node.

[0047] In shared nothing systems, when a data item is transferred into cache in dynamic memory from a secondary data resource in the form of a static storage device, the node that has direct access to the static storage device reads the data item into the dynamic memory of the node. If a cache in another dynamic memory of another node is the primary cache for the data item, then that data item will be transferred there, resulting in an additional transfer of the data item. This may be avoided by treating the node that has direct access to the secondary resource that holds the data item as the primary cache for the data item.

HARDWARE OVERVIEW

[0048] Figure 4 is a block diagram that illustrates a computer system 400 which may be used to implement an embodiment of the invention. Computer system 400 includes a bus 402 or other communication mechanism for communicating information, and a processor 404 cou-

pled with bus 402 for processing information. Computer system 400 also includes a main memory 406, such as a random access memory (RAM) or other dynamic storage device, coupled to bus 402 for storing information and instructions to be executed by processor 404. Main memory 406 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 404. Computer system 400 further includes a read only memory (ROM) 408 or other static storage device coupled to bus 402 for storing static information and instructions for processor 404. A storage device 410, such as a magnetic disk or optical disk, is provided and coupled to bus 402 for storing information and instructions.

[0049] Computer system 400 may be coupled via bus 402 to a display 412, such as a cathode ray tube (CRT), for displaying information to a computer user. An input device 414, including alphanumeric and other keys, is coupled to bus 402 for communicating information and command selections to processor 404. Another type of user input device is cursor control 416, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor 404 and for controlling cursor movement on display 412. This input device typically has two degrees of freedom in two axes, a first axis (e.g., x) and a second axis (e.g., y), that allows the device to specify positions in a plane.

[0050] The invention is related to the use of computer system 400 for implementing the techniques described herein. According to one embodiment of the invention, those techniques are implemented by computer system 400 in response to processor 404 executing one or more sequences of one or more instructions contained in main memory 406. Such instructions may be read into main memory 406 from another computer-readable medium, such as storage device 410. Execution of the sequences of instructions contained in main memory 406 causes processor 404 to perform the process steps described herein. In alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions to implement the invention. Thus, embodiments of the invention are not limited to any specific combination of hardware circuitry and software.

[0051] The term "computer-readable medium" as used herein refers to any medium that participates in providing instructions to processor 404 for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media includes, for example, optical or magnetic disks, such as storage device 410. Volatile media includes dynamic memory, such as main memory 406. Transmission media includes coaxial cables, copper wire and fiber optics, including the wires that comprise bus 402. Transmission media can also take the form of acoustic or light waves, such as those generated during radio-wave and infra-red data communications.

[0052] Common forms of computer-readable media in-

clude, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punchcards, papertape, any other physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave as described hereinafter, or any other medium from which a computer can read.

[0053] Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to processor 404 for execution. For example, the instructions may initially be carried on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to computer system 400 can receive the data on the telephone line and use an infra-red transmitter to convert the data to an infra-red signal. An infra-red detector can receive the data carried in the infra-red signal and appropriate circuitry can place the data on bus 402. Bus 402 carries the data to main memory 406, from which processor 404 retrieves and executes the instructions. The instructions received by main memory 406 may optionally be stored on storage device 410 either before or after execution by processor 404.

[0054] Computer system 400 also includes a communication interface 418 coupled to bus 402. Communication interface 418 provides a two-way data communication coupling to a network link 420 that is connected to a local network 422. For example, communication interface 418 may be an integrated services digital network (ISDN) card or a modem to provide a data communication connection to a corresponding type of telephone line. As another example, communication interface 418 may be a local area network (LAN) card to provide a data communication connection to a compatible LAN. Wireless links may also be implemented. In any such implementation, communication interface 418 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

[0055] Network link 420 typically provides data communication through one or more networks to other data devices. For example, network link 420 may provide a connection through local network 422 to a host computer 424 or to data equipment operated by an Internet Service Provider (ISP) 426. ISP 426 in turn provides data communication services through the world wide packet data communication network now commonly referred to as the "Internet" 428. Local network 422 and Internet 428 both use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on network link 420 and through communication interface 418, which carry the digital data to and from computer system 400, are exemplary forms of carrier waves transporting the information.

[0056] Computer system 400 can send messages and receive data, including program code, through the net-

work(s), network link 420 and communication interface 418. In the Internet example, a server 430 might transmit a requested code for an application program through Internet 428, ISP 426, local network 422 and communication interface 418. In accordance with the invention, one such downloaded application implements the techniques described herein.

[0057] The received code may be executed by processor 404 as it is received, and/or stored in storage device 410, or other non-volatile storage for later execution. In this manner, computer system 400 may obtain application code in the form of a carrier wave.

[0058] In the foregoing specification, the invention has been described with reference to specific embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the scope of the invention. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

Claims

1. A method for managing replacement of data stored in buffers of a plurality of caches (114, 134, 154) that each reside on a node (110, 130, 150) of a set of one or more nodes (110, 130, 150), wherein each node of said set of one or more nodes includes one or more caches of said plurality of caches, wherein the plurality of caches includes a first cache, the method comprising the steps of:

from a plurality of buffers available for replacement in the first cache (114), selecting for replacement a first buffer from said first cache (114) based on one or more factors, wherein said first buffer is currently used to store a first data item, wherein said one or more factors include either:

the state of at least one other cache (134, 154) of said plurality of caches, or
the configuration of said at least one other cache (134, 154) of said plurality of caches;

replacing said first data item stored in said first buffer with a second data item;
transferring said first data item to another particular cache (134, 154) of said plurality of caches;
in response to transferring said first data item to the particular cache (134, 154), generating certain data that indicates that said first buffer contains data that may be removed from said first cache (114); and
wherein said one or more factors include that said certain data indicates that said first buffer contains data that may be replaced.

2. The method of claim 1, wherein said one or more factors includes whether said first data item resides in the at least one other cache (134, 154) of said plurality of caches.

3. The method of claim 1, wherein the step of generating certain data is not performed when said first data item belongs to a particular subset of data items stored in the first cache (114).

4. The method of claim 3, wherein said particular subset of data items include the most frequently accessed set of data items that currently reside in said first cache (114).

5. The method of claim 1, wherein:

before replacing said second data item, said first buffer is associated with a first cache retention value and a second buffer is associated with a second cache retention value, wherein a second cache (134) includes said second buffer; and wherein said one or more factors include that said first cache retention value is less than said second cache retention value.

6. The method of claim 5, wherein where said first cache (114) is managed by a first cache manager (112) and said second cache (134) is managed by second cache manger (132), wherein the method further includes the steps of:

said first cache manager (112) receiving on behalf of said second cache manager (132) a request for a direct transfer of a particular data item stored in the first cache (114);
in response to said first cache manager (112) receiving said request, said first cache manager (112) transmitting said first cache retention value and said particular data item to said second cache (134), wherein said first cache retention value indicates a lowest cache retention value of at least a subset of buffers in said first cache (114).

7. The method of claim 5, further including the steps of:

said first cache (114) receiving said second data item transmitted from said second cache (134) in response to said second cache manager (132) determining that said second cache retention value is greater than said first cache retention value; and
wherein said factors include that said first cache received said second data item transmitted from said second cache in response to said second cache manager (132) determining that said second cache retention value is greater than said

first cache retention value.

8. The method of claim 1, further including the step of transferring a copy of said first data item to a second cache (134) of a second node (130) without first durably storing said first data item from said first cache (114) to persistent storage. 5
9. The method of claim 1, wherein said plurality of caches (114, 134, 154) are each managed by a database system. 10
10. The method of claim 1, wherein said one or more factors include whether said data item becomes eligible for replacement when transferred from another particular cache (134, 154) of said plurality of caches. 15
11. A computer-readable medium carrying one or more sequences of instructions which, when executed by one or more processors, causes the one or more processors to perform the method recited in one of the claims 1-10. 20

Revendications

1. Procédé de gestion du remplacement de données stockées dans des tampons d'une pluralité de caches (114, 134, 154) qui résident chacun sur un noeud (110, 130, 150) d'un ensemble d'un ou plusieurs noeuds (110, 130, 150), dans lequel chaque noeud dudit ensemble d'un ou plusieurs noeuds comprend un ou plusieurs caches de ladite pluralité de caches, dans lequel la pluralité de caches comprend un premier cache, le procédé comprenant les étapes consistant à : 30
 - à partir d'une pluralité de tampons disponibles pour le remplacement dans le premier cache (114), sélectionner pour le remplacement un premier tampon à partir dudit premier cache (114) en fonction d'un ou plusieurs facteurs, dans lequel ledit premier tampon est actuellement utilisé pour stocker un premier élément de données, dans lequel lesdits un ou plusieurs facteurs comprennent : 35
 - l'état d'au moins un autre cache (134, 154) de ladite pluralité de caches, ou 40
 - la configuration dudit au moins un autre cache (134, 154) de ladite pluralité de caches ; 45
 - remplacer ledit premier élément de données stocké dans ledit premier tampon par un deuxième élément de données ; 50
 - transférer ledit premier élément de données vers un autre cache particulier (134, 154) de la-

dite pluralité de caches ;
 en réponse au transfert dudit premier élément de données vers le cache particulier (134, 154), générer certaines données qui indiquent que ledit premier tampon contient des données qui peuvent être supprimées dudit premier cache (114) ; et
 dans lequel lesdits un ou plusieurs facteurs comprennent que lesdites certaines données indiquent que ledit premier tampon contient des données qui peuvent être remplacées.

2. Procédé selon la revendication 1, dans lequel lesdits un ou plusieurs facteurs comprennent si ledit premier élément de données réside dans ledit au moins un autre cache (134, 154) de ladite pluralité de caches.
3. Procédé selon la revendication 1, dans lequel l'étape consistant à générer certaines données n'est pas réalisée lorsque ledit premier élément de données appartient à un sous-ensemble particulier d'éléments de données stocké dans le premier cache (114).
4. Procédé selon la revendication 3, dans lequel ledit sous-ensemble particulier d'éléments de données comprend l'ensemble d'éléments de données accédé le plus fréquemment qui réside actuellement dans ledit premier cache (114). 25
5. Procédé selon la revendication 1, dans lequel : 30
 - avant de remplacer ledit deuxième élément de données, ledit premier tampon est associé à une première valeur de rétention de cache et un deuxième tampon est associé à une deuxième valeur de rétention de cache, dans lequel un deuxième cache (134) comprend ledit deuxième tampon ; et 35
 - dans lequel lesdits un ou plusieurs facteurs comprennent que ladite première valeur de rétention de cache est inférieure à ladite deuxième valeur de rétention de cache.
6. Procédé selon la revendication 5, dans lequel ledit premier cache (114) est géré par un premier gestionnaire de cache (112) et ledit deuxième cache (134) est géré par un deuxième gestionnaire de cache (132), dans lequel le procédé comprend en outre les étapes consistant en ce que : 45

ledit premier gestionnaire de cache (112) reçoit de la part dudit deuxième gestionnaire de cache (132) une demande pour un transfert direct d'un élément de données particulier stocké dans le premier cache (114) ;
 en réponse audit premier gestionnaire de cache (112) qui reçoit ladite demande, ledit premier

- gestionnaire de cache (112) transmet ladite première valeur de rétention de cache et ledit élément de données particulier vers ledit deuxième cache (134), dans lequel ladite première valeur de rétention de cache indique une valeur de rétention de cache la plus basse d'au moins un sous-ensemble de tampons dans ledit premier cache (114). 5
7. Procédé selon la revendication 5, comprenant en outre les étapes consistant en ce que : 10
- ledit premier cache (114) reçoit ledit deuxième élément de données émis depuis ledit deuxième cache (134) en réponse audit deuxième gestionnaire de cache (132) déterminant que ladite deuxième valeur de rétention de cache est supérieure à ladite première valeur de rétention de cache ; et 15
- dans lequel lesdits facteurs comprennent que ledit premier cache a reçu ledit deuxième élément de données émis depuis ledit deuxième cache en réponse audit deuxième gestionnaire de cache (132) déterminant que ladite valeur de rétention de cache est supérieure à ladite première valeur de rétention de cache. 20 25
8. Procédé selon la revendication 1, comprenant en outre l'étape consistant à transférer une copie dudit premier élément de données vers un deuxième cache (134) d'un deuxième noeud (130) sans tout d'abord stocker durablement ledit premier élément de données provenant dudit premier cache (114) vers un stockage persistant. 30 35
9. Procédé selon la revendication 1, dans lequel chacun de ladite pluralité de caches (114, 134, 154) est géré par un système de base de données. 40
10. Procédé selon la revendication 1, dans lequel lesdits un ou plusieurs facteurs comprennent si ledit élément de données devient éligible pour le remplacement lorsqu'il est transféré depuis un autre cache particulier (134, 154) de ladite pluralité de caches. 45
11. Support lisible par ordinateur portant une ou plusieurs séquences d'instructions qui, lorsqu'elles sont exécutées par un ou plusieurs processeurs, amène lesdits un ou plusieurs processeurs à réaliser le procédé selon l'une quelconque des revendications 1 à 10. 50
- Patentansprüche** 55
1. Verfahren zum Verwalten des Ersetzens von Datenwerten, die in Puffern einer Mehrzahl von Zwischen-
- speichern (114, 134, 154) gespeichert sind, die jeweils auf einem Knoten (110, 130, 150) eines Satzes von einem oder einer Mehrzahl von Knoten (110, 130, 150) angeordnet sind, wobei jeder Knoten des Satzes von einem oder einer Mehrzahl von Knoten einen oder eine Mehrzahl von Zwischenspeichern der Mehrzahl von Zwischenspeichern aufweist, wobei die Mehrzahl von Zwischenspeichern einen ersten Zwischenspeicher aufweist, wobei das Verfahren die Schritte aufweist:
- von einer Mehrzahl von Puffern, die zum Ersetzen in dem ersten Zwischenspeicher (114) verfügbar sind, Auswählen zum Ersetzen eines ersten Puffers aus dem ersten Zwischenspeicher (114) auf der Basis von einem oder einer Mehrzahl von Faktoren, wobei der erste Puffer gegenwärtig zum Speichern eines ersten Datenwerts verwendet wird, wobei der eine oder die Mehrzahl von Faktoren entweder aufweisen:
- den Zustand von wenigstens einem anderen Zwischenspeicher (134, 154) der Mehrzahl von Zwischenspeichern, oder
- die Konfiguration des wenigstens einen anderen Zwischenspeichers (134, 154) der Mehrzahl von Zwischenspeichern;
- Ersetzen des ersten Datenwerts, der in dem ersten Puffer gespeichert ist, durch einen zweiten Datenwert;
- Übertragen des ersten Datenwerts in einen anderen bestimmten Zwischenspeicher (134, 154) der Mehrzahl von Zwischenspeichern;
- Erzeugen bestimmter Daten, die anzeigen, dass der erste Puffer Daten aufweist, die aus dem ersten Zwischenspeicher (114) entfernt werden können, in Antwort auf das Übertragen des ersten Datenwerts in den bestimmten Zwischenspeicher (134, 154); und
- wobei der eine oder die Mehrzahl von Faktoren aufweisen, dass die bestimmten Daten anzeigen, dass der erste Puffer einen Datenwert aufweist, der ersetzt werden kann.
2. Verfahren gemäß Anspruch 1, wobei der eine oder die Mehrzahl von Faktoren aufweisen, ob der erste Datenwert in dem wenigstens einem anderen Zwischenspeicher (134, 154) der Mehrzahl von Zwischenspeichern angeordnet ist.
3. Verfahren gemäß Anspruch 1, wobei der Schritt des Erzeugens bestimmter Daten nicht durchgeführt wird, wenn der erste Datenwert zu einer bestimmten Untermenge von Datenwerten gehört, die in dem ersten Zwischenspeicher (114) gespeichert ist.
4. Verfahren gemäß Anspruch 3, wobei die bestimmte

Untermenge von Datenwerten den Satz von Datenwerten aufweist, auf die am häufigsten zugegriffen wird, und der gegenwärtig in dem ersten Zwischenspeicher (114) angeordnet ist.

5. Verfahren gemäß Anspruch 1, wobei:

vor dem Ersetzen des zweiten Datenwerts dem ersten Puffer ein erster Zwischenspeicher-Beibehaltungswert zugeordnet wird, und einem zweiten Puffer ein zweiter Zwischenspeicher-Beibehaltungswert zugeordnet wird, wobei ein zweiter Zwischenspeicher (134) den zweiten Puffer aufweist; und

wobei der eine oder die Mehrzahl von Faktoren aufweisen, dass der erste Zwischenspeicher-Beibehaltungswert kleiner ist als der zweite Zwischenspeicher-Beibehaltungswert.

6. Verfahren gemäß Anspruch 5, wobei dort, wo der erste Zwischenspeicher (114) von einem ersten Zwischenspeicher-Manager (112) verwaltet wird und der zweite Zwischenspeicher (134) von einem zweiten Zwischenspeicher-Manager (132) verwaltet wird, das Verfahren ferner die Schritte aufweist:

der erste Zwischenspeicher-Manager (112) empfängt für den zweiten Zwischenspeicher-Manager (132) eine Anfrage für eine direkte Übertragung eines bestimmten Datenwerts, die in dem ersten Zwischenspeicher (114) gespeichert ist;

in Antwort auf den ersten Zwischenspeicher-Manager (112), der die Anfrage empfängt, überträgt der erste Zwischenspeicher-Manager (112) den ersten Zwischenspeicher-Beibehaltungswert und den bestimmten Datenwert zu dem zweiten Zwischenspeicher (134), wobei der erste Zwischenspeicher-Beibehaltungswert einen kleinsten Zwischenspeicher-Beibehaltungswert aus wenigstens einer Untermenge von Puffern in dem ersten Zwischenspeicher (114) anzeigt.

7. Verfahren gemäß Anspruch 5, das ferner die Schritte aufweist:

der erste Zwischenspeicher (114), der den zweiten Datenwert empfängt, der von dem zweiten Zwischenspeicher (134) in Antwort auf den zweiten Zwischenspeicher-Manager (132) übertragen wird, die bestimmt, dass der zweite Zwischenspeicher-Beibehaltungswert größer ist als der erste Zwischenspeicher-Beibehaltungswert; und

wobei die Faktoren aufweisen, dass der erste Zwischenspeicher den zweiten Datenwert empfängt, der von dem zweiten Zwischenspeicher

in Antwort auf den zweiten Zwischenspeicher-Manager (132) übertragen wird, die bestimmt, dass der zweite Zwischenspeicher-Beibehaltungswert größer ist als der erste Zwischenspeicher-Beibehaltungswert.

8. Verfahren gemäß Anspruch 1, ferner aufweisend den Schritt des Übertragens einer Kopie des ersten Datenwerts zu einem zweiten Zwischenspeicher (134) eines zweiten Knotens (130), ohne zuerst den ersten Datenwert von dem ersten Zwischenspeicher (114) in einem persistenten Speicher dauerhaft zu speichern.

9. Verfahren gemäß Anspruch 1, wobei die Mehrzahl von zwischenspeichern (114, 134, 154) jeweils von einem Datenbanksystem verwaltet wird.

10. Verfahren gemäß Anspruch 1, wobei der eine oder die Mehrzahl von Faktoren aufweisen, ob der Datenwert zum Ersetzen geeignet sein wird, wenn er von einem anderen bestimmten Zwischenspeicher (134, 154) der Mehrzahl von Zwischenspeicher übertragen wird.

11. Computer-lesbares Medium, das eine oder eine Mehrzahl von Befehlssequenzen trägt, die, wenn sie von einem oder einer Mehrzahl von Prozessoren ausgeführt werden, bewirken, dass der eine oder die Mehrzahl von Prozessoren das Verfahren ausführen, das in einem der Ansprüche 1-10 aufgezählt ist.

FIG. 1

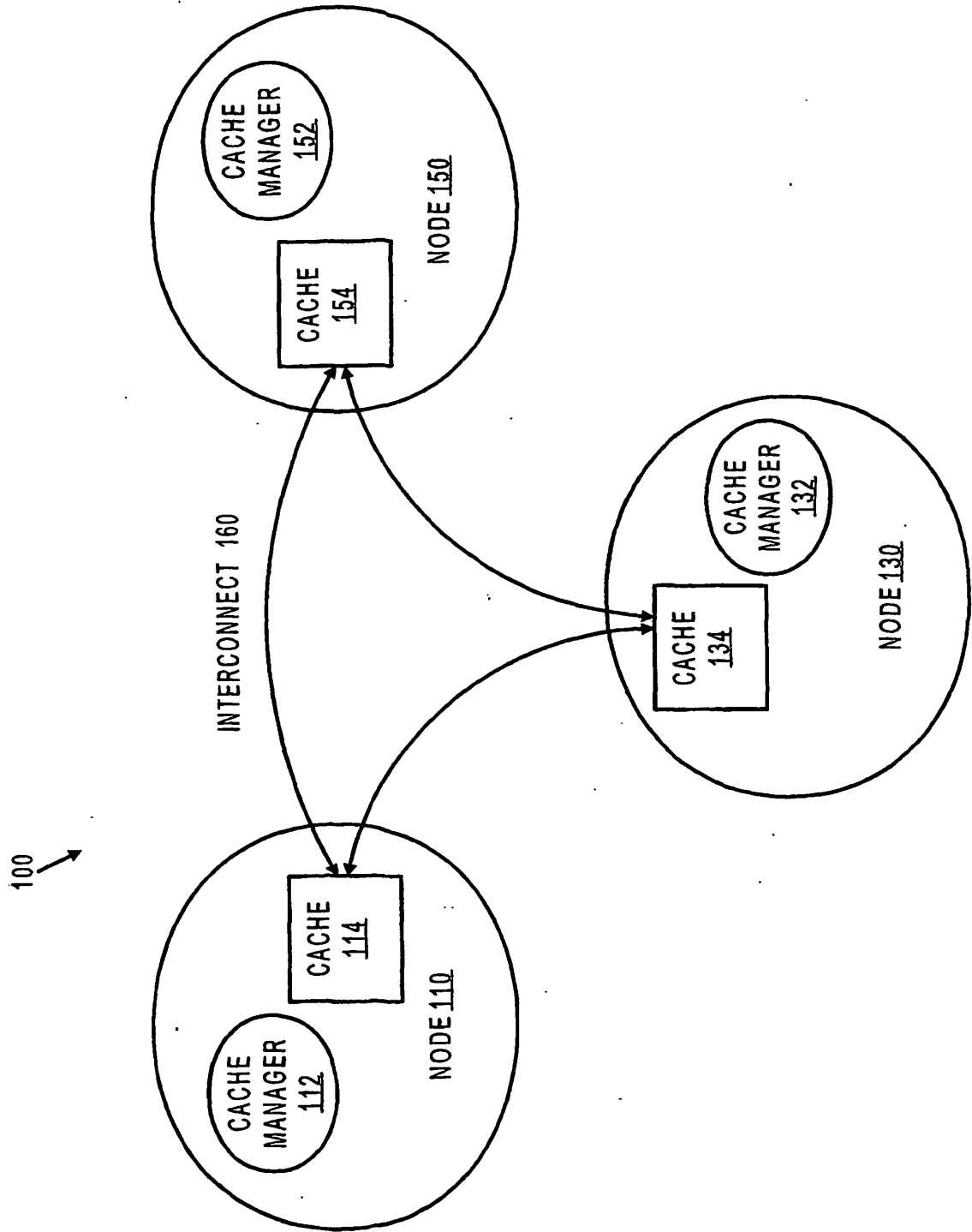


FIG. 2

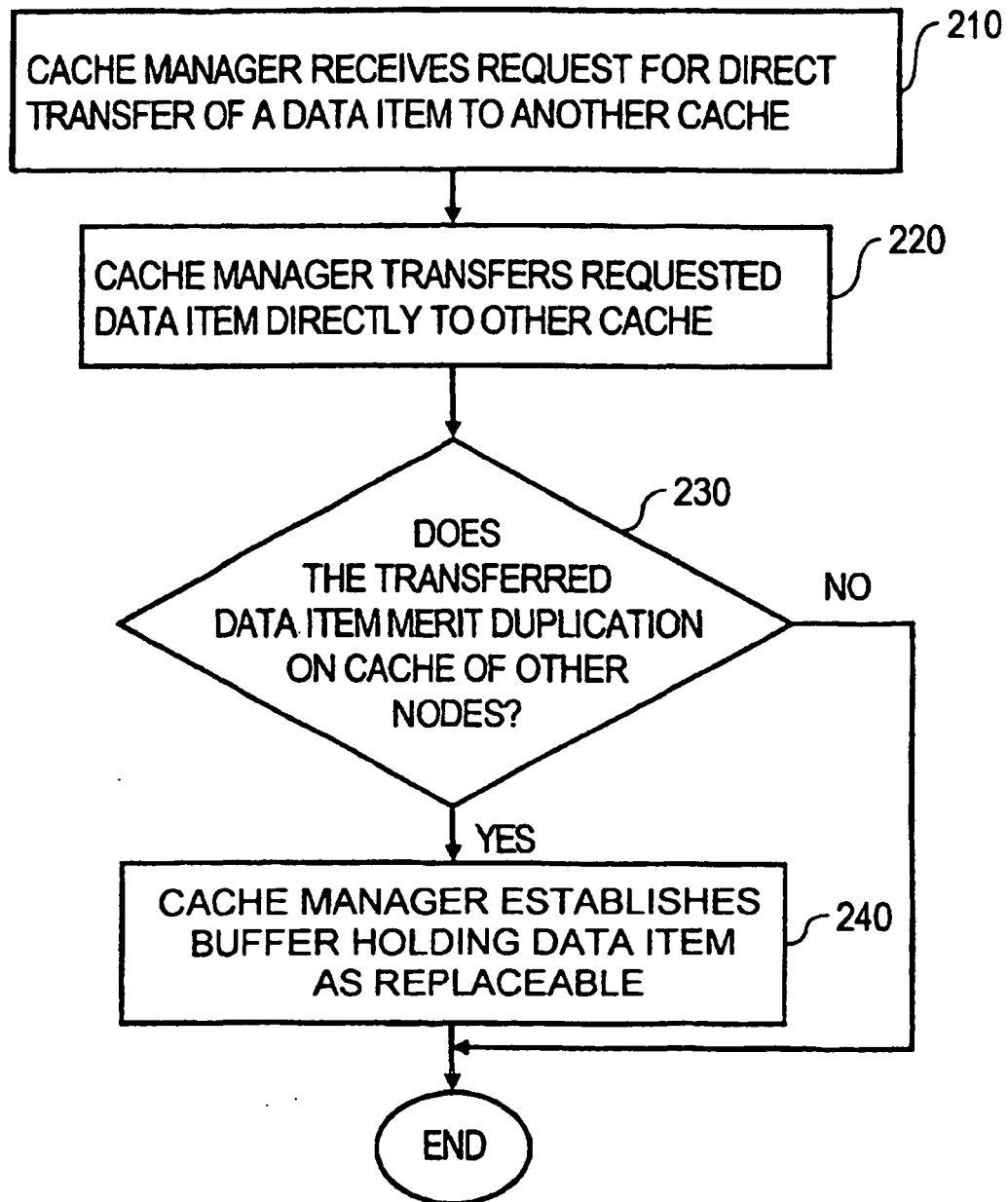


FIG. 3

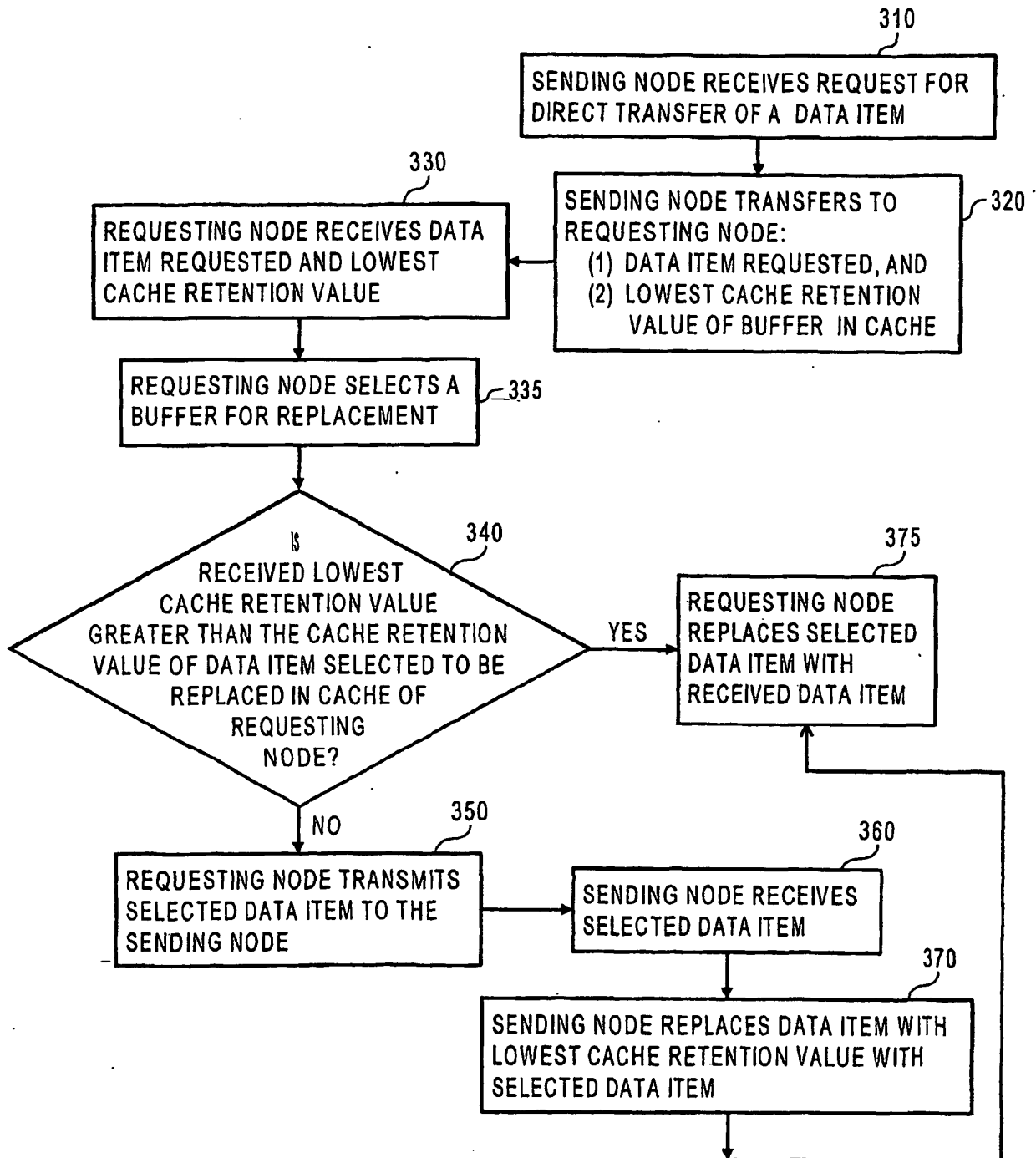


FIG. 4

